Characterizing Strain Variation in Engineered E. coli Using a Multi-Omics-Based Workflow

Brunk, Elizabeth; George, Kevin W.; Alonso-Gutierrez, Jorge; Tjompson, Mitchell; Baidoo, Edward; Wang, George; Petzold, Christopher J.; McCloskey , Douglas; Monk, Jonathan; Yang, Laurence

Total number of authors:
18

Published in:
Cell Systems

Link to article, DOI:
10.1016/j.cels.2016.04.004

Publication date:
2016

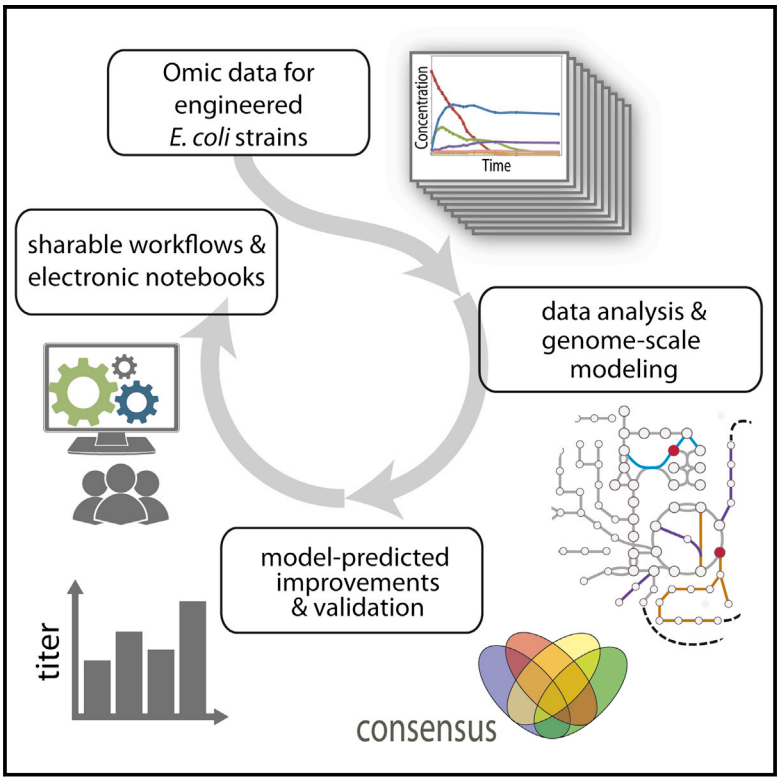
Document Version
Publisher's PDF, also known as Version of record

Article

Cell Systems

Characterizing Strain Variation in Engineered *E. coli* Using a Multi-Omics-Based Workflow

Graphical Abstract



Authors

Elizabeth Brunk, Kevin W. George, Jorge Alonso-Gutierrez, ..., Jay D. Keasling, Bernhard O. Palsson, Taek Soon Lee

Correspondence

palsson@eng.ucsd.edu (B.O.P.), tslee@lbl.gov (T.S.L.)

In Brief

Brunk et al. develop a workflow to assess and interpret multi-omics data and use it to characterize strain variation in biofuel-producing *E. coli*.

Highlights

- Eight biofuel-producing *E. coli* strains are assessed with multi-omics data
- Three-stage workflow incorporates computational, systems, and synthetic biology
- Interactions between synthetic and endogenous metabolic pathways are explored
- Genome-scale modeling identifies a knockout that increases target production

Brunk et al., 2016, Cell Systems *2*, 335–346
May 25, 2016
http://dx.doi.org/10.1016/j.cels.2016.04.004

Cell Systems
Article

Characterizing Strain Variation in Engineered *E. coli* Using a Multi-Omics-Based Workflow

Elizabeth Brunk,[1,2,3,9] Kevin W. George,[1,3,9,10] Jorge Alonso-Gutierrez,[1,3] Mitchell Thompson,[1,4] Edward Baidoo,[1,3] George Wang,[1,3] Christopher J. Petzold,[1,3] Douglas McCloskey,[2] Jonathan Monk,[2] Laurence Yang,[2] Edward J. O'Brien,[2] Tanveer S. Batth,[1] Hector Garcia Martin,[1,3] Adam Feist,[2,3] Paul D. Adams,[1,6] Jay D. Keasling,[1,3,5,7,8] Bernhard O. Palsson,[2,5,*] and Taek Soon Lee[1,3,*]

[1]Joint Bioenergy Institute (JBEI), 5885 Hollis Street, Emeryville, CA 94608, USA
[2]Department of Bioengineering, University of California, San Diego, San Diego, CA 92093, USA
[3]Biological Systems & Engineering Division, Lawrence Berkeley National Laboratory, Berkeley, CA 94720, USA
[4]Department of Plant and Microbial Biology, University of California, Berkeley, Berkeley, CA 94720, USA
[5]The Novo Nordisk Foundation Center for Biosustainability, Technical University of Denmark, 2970 Horsholm, Denmark
[6]Molecular Biophysics and Integrated Bioimaging Division, Lawrence Berkeley National Laboratory, Berkeley, CA 94720, USA
[7]Department of Chemical & Biomolecular Engineering, University of California, Berkeley, Berkeley, CA 94720, USA
[8]Department of Bioengineering, University of California, Berkeley, Berkeley, CA 94720, USA
[9]Co-first author
[10]Present address: Amyris, 5885 Hollis Street, Emeryville, CA 94608, USA
*Correspondence: palsson@eng.ucsd.edu (B.O.P.), tslee@lbl.gov (T.S.L.)
http://dx.doi.org/10.1016/j.cels.2016.04.004

SUMMARY

Understanding the complex interactions that occur between heterologous and native biochemical pathways represents a major challenge in metabolic engineering and synthetic biology. We present a workflow that integrates metabolomics, proteomics, and genome-scale models of *Escherichia coli* metabolism to study the effects of introducing a heterologous pathway into a microbial host. This workflow incorporates complementary approaches from computational systems biology, metabolic engineering, and synthetic biology; provides molecular insight into how the host organism microenvironment changes due to pathway engineering; and demonstrates how biological mechanisms underlying strain variation can be exploited as an engineering strategy to increase product yield. As a proof of concept, we present the analysis of eight engineered strains producing three biofuels: isopentenol, limonene, and bisabolene. Application of this workflow identified the roles of candidate genes, pathways, and biochemical reactions in observed experimental phenomena and facilitated the construction of a mutant strain with improved productivity. The contributed workflow is available as an open-source tool in the form of iPython notebooks.

INTRODUCTION

The confluence of high-throughput omics technologies and quantitative systems biology has dramatically enhanced our ability to probe biological phenomena across a vast range of chemical and biological scales (de Jong et al., 2012; Kuehnbaum and Britz-McKibbin, 2013; Tyo et al., 2007). Large-scale improvements in data coverage and measurement fidelity enable the quantitative tracking of dynamic changes in RNA transcripts, ribosome profiling, proteins, and metabolites in unprecedented detail (Fuhrer and Zamboni, 2015; Gross, 2011; Kahn, 2011; Metzker, 2010; Zhang et al., 2014). Yet, current computational tools for handling such data are rapidly becoming inadequate when compared to the amount of omics data that can now be generated (Stephens et al., 2015). This challenge, referred to as "Big Data to Knowledge" (Margolis et al., 2014), requires balancing the deluge of experimental "big data" with a solid, theoretical basis for its interpretation.

Impediments to realizing the potential impact of big data resources include a lack of appropriate in silico tools, poor data accessibility, and insufficient cross-disciplinary training. Current computational methods are limited in their capacity to accommodate an increasingly diverse range of experimental techniques and contextualize new data within existing datasets (Berger et al., 2013). Furthermore, the skillsets required of scientists in the era of big data now extend beyond the traditional scope of biochemistry and molecular biology to include bioinformatics, biostatistics, and computer science. Hence, despite the interest to collaborate or use tools from an orthogonal field of research, domain-specific jargon is yet another obstacle to overcome by the prospective practitioner in big data science (Rolfsson and Palsson, 2015).

In this work, we hope to lower the barrier of entry into computational systems biology by creating a framework upon which disparate biological data types can be analyzed and interpreted. We take advantage of three synergistic, accelerating domains of science—systems biology, metabolic engineering, and synthetic biology—to develop a workflow that reconciles systems-level, multi-omics analysis, and genome-scale modeling with synthetic pathway engineering. While the collection of targeted omics data has supported a number of metabolic engineering efforts (Alonso-Gutierrez et al., 2015; George et al., 2014; Han et al., 2001, 2003; Kabir and Shimizu, 2003; Landels et al., 2015; Lee

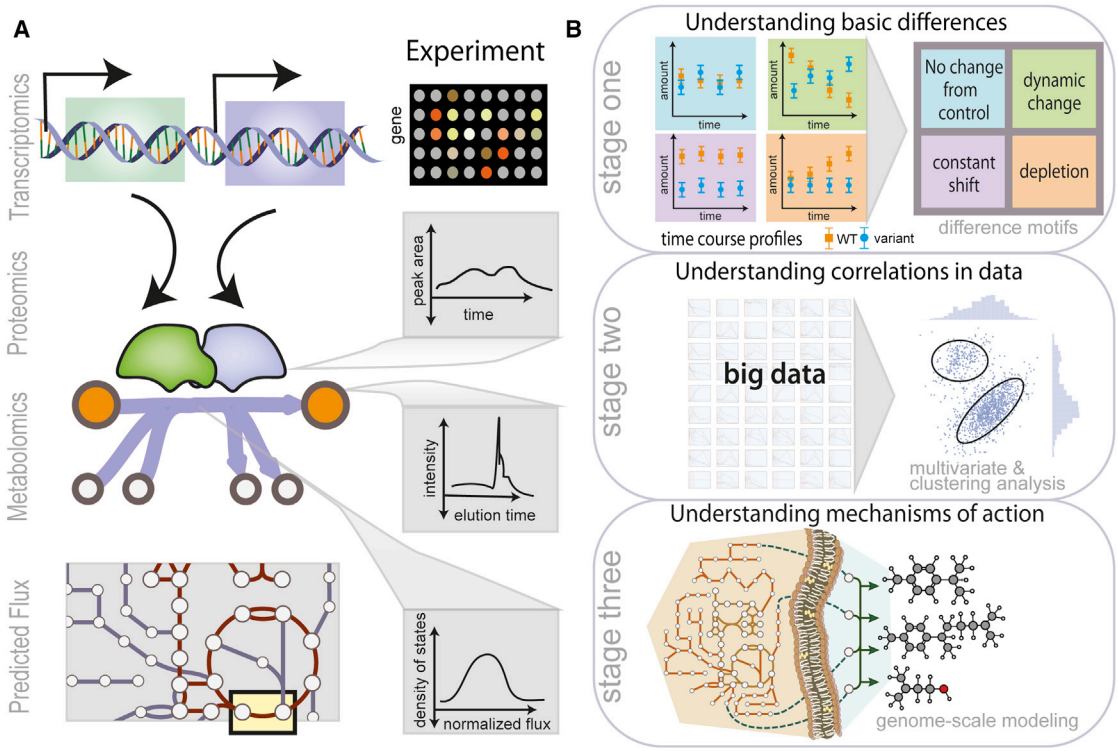


Figure 1. A Workflow for Bridging the Genotype-Phenotype Relationship with Multi-omics Data and Genome-Scale Models of *E. coli* Metabolism Expressing Heterologous Pathways

(A) Multi-scale data types that are generally collected to elucidate changes in metabolic phenotypes of different strains.

(B) Our workflow involves a hierarchical staging of computational analysis methods: (1) basic strain differences; (2) relevant patterns and correlations in the data; (3) mechanisms of action in the context of a genome-scale network that can explain apparent differences in strain behavior.

et al., 2005), the extraction of biologically meaningful information from highly dimensional multi-omics datasets remains a continual challenge (Kwok, 2010; Nielsen et al., 2014; Palsson and Zengler, 2010). Engineering strategies such as the design-build-test-analyze (DBTA) cycle (Bailey, 1991) attempt to side step this issue through rapid iteration and strain assessment, but the "analyze" phase of the cycle is often limited by a narrow focus on one or two experimental outputs such as product titer. This motivates the development of tools to better characterize the biological components of these complex systems, decrease the heavy reliance on iterative trial-and-error, and bring biological engineering closer to other, more rational, engineering disciplines.

To address this multi-layered challenge, our hierarchical workflow consists of three stages (Figure 1). In the first stage, basic strain differences are assessed through a global analysis of computationally derived "dynamic difference profiles." The second stage uses multivariate analysis to identify relevant patterns and correlations in key metabolites and proteins. In the last stage, these inputs are reconciled with genome-scale models to identify perturbed metabolic nodes that are subsequently validated and investigated as engineering leads. We apply this framework to eight engineered strains of *E. coli* producing three isoprenoid-derived advanced biofuels and demonstrate that this strategy is capable of clarifying convoluted metabolic network responses, identifying potential bottlenecks, and elucidating the complex interplay between synthetic and endogenous *E. coli* metabolism.

RESULTS AND DISCUSSION

Pathway Description, Strain Selection, and Multi-omics Data Generation

In synthetic biology, the design of efficient "cell factories" typically involves the introduction of heterologous genes and metabolic pathways into a microbial host. In the last decade, broad classes of chemicals including isoprenoids, polyketides, branched chain alcohols, and fatty acids have been successfully produced using a variety of microbial hosts and renewable, bio-based materials (Jullesson et al., 2015; Peralta-Yahya et al., 2012). The native mevalonate pathway from *Saccharomyces cerevisiae*, which consists of six reactions that convert acetyl-CoA into the isoprenoid precursor isopentenyl diphosphate (ipdp or IPP), has been heterologously expressed in *E. coli* (Martin et al., 2003) and adapted to produce a variety of terpene fuels and chemicals (George et al., 2015a). By expressing additional genes, this core pathway has been modified to produce C_5 (hemiterpene) isopentenol (Chou and Keasling, 2012), C_{10} (monoterpene) limonene (Alonso-Gutierrez et al., 2013), and C_{15} (sesquiterpene) bisabolene (Peralta-Yahya et al., 2011), terpenes that serve as drop-in replacements for gasoline, jet fuel, and diesel, respectively.

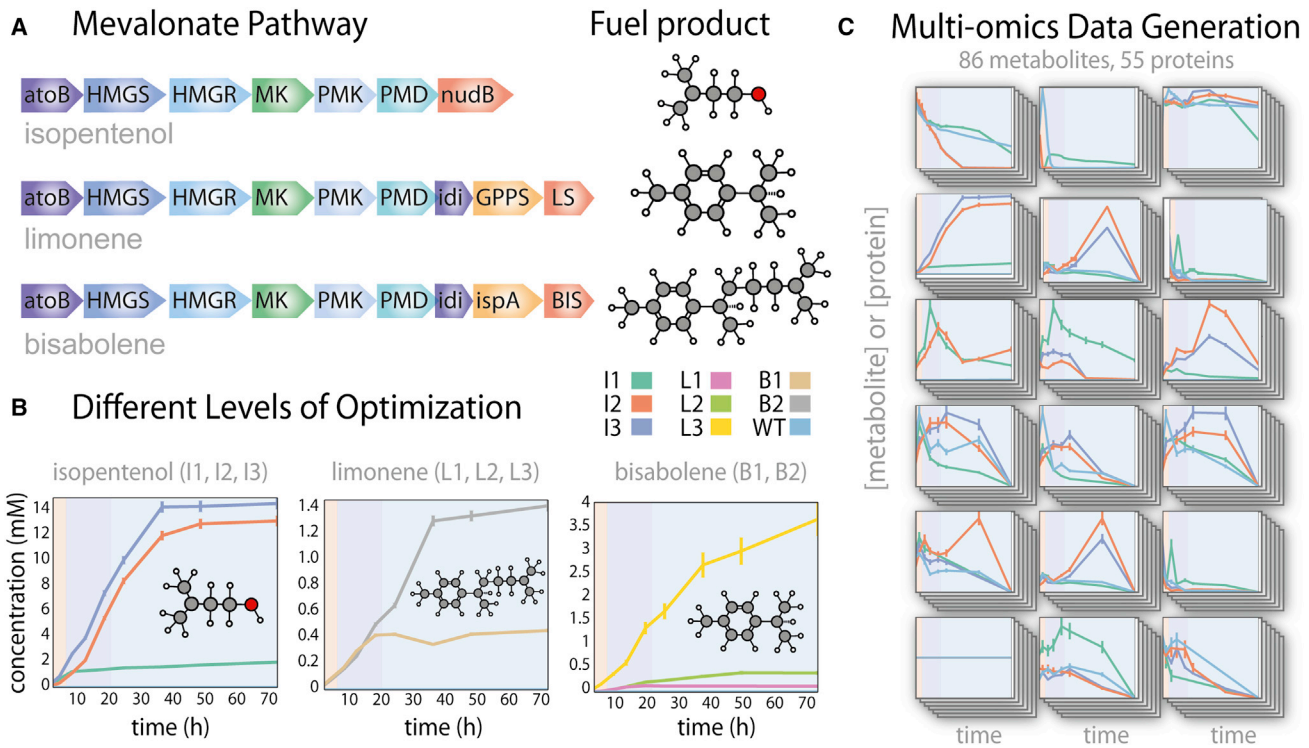


Figure 2. Pathway Assembly, Strain Selection, and Multi-omics Data Generation

(A) This study characterizes three versions of a heterologous mevalonate pathway engineered to synthesize isopentenol, limonene, and bisabolene.

(B) Over a 72-hr time course, the engineered strains show various levels of fuel production due to changes in heterologous pathway architecture and expression. Each strain is indicated by its respective color, shown in the legend to the right.

(C) The nine strains were further analyzed using a multi-omics approach during batch fermentation to generate detailed omics profiles.

Optimization of each of these heterologous pathways has yielded strains with significantly improved titers through methods such as codon optimization of poorly expressed genes, promoter supplementation, altered operon order, and changes in plasmid copy number (Alonso-Gutierrez et al., 2015; George et al., 2014, 2015b; Peralta-Yahya et al., 2011). Though the titer of each fuel target has consistently improved, the impact of these optimizations on endogenous *E. coli* metabolism has yet to be comprehensively explored (Figure S1). Given that previous strain optimization has focused primarily on the mevalonate pathway itself, we suspected that a systematic exploration of the interplay between heterologous pathway engineering and endogenous metabolism could better characterize strain variation, identify perturbed metabolic nodes, and ultimately yield new engineering targets. To explore this issue, we selected representative strains for each biofuel (Figure 2A) with different levels of optimization (Figure 2B) and collected extensive omics data (Figure 2C) for both heterologous and endogenous metabolism in a fermentation time course.

Our analysis included three isopentenol-producing strains (I1–I3), three limonene-producing strains (L1–L3), two bisabolene-producing strains (B1 and B2), and wild-type *E. coli* DH1 (WT) (nine strains total; Figure S2 and Table S1). The numbering of the strains in each set represents their overall performance (product yield) and evolution of the optimization process (i.e., "1" represents non-optimized pathway and "2" or "3" represents variants with better performance). Samples were collected to measure cell growth, product titer, intracellular and extracellular metabolites, and selected proteins at multiple time-points (0–72 hr post-induction) in the batch fermentation. Altogether, our analysis included the absolute quantification of more than 80 metabolites and the relative quantification (via a targeted SRM method [Picotti and Aebersold, 2012]) of more than 50 proteins or protein complexes spanning key "nodes" in heterologous (i.e., mevalonate pathway) and endogenous metabolism (Data S1 and S2).

Stage One: Integrating Multi-omics Data and Profiling Batch Fermentation Dynamics

Stage one of the workflow (Figure 3) integrates raw, multi-dimensional omics data to identify basic differences between strains. First, we assign test (e.g., engineered strain) and control (e.g., WT) conditions, which can vary depending upon the question being addressed. We take the difference of measured metabolite concentrations in the test and control conditions at each time point to "bin" the pairwise differences into one of six "dynamic difference profiles" that describe the behavior of the test condition relative to the control (e.g., "no change," "constant," "deviation," "return," "shift," and "transient"; Figure 3). With this framework in place, thousands of omics inputs can be rapidly "filtered" into distinct profiles to facilitate large-scale strain comparisons and statistical analysis.

We generated dynamic difference profiles for the eight biofuel-producing strains to examine which metabolites were

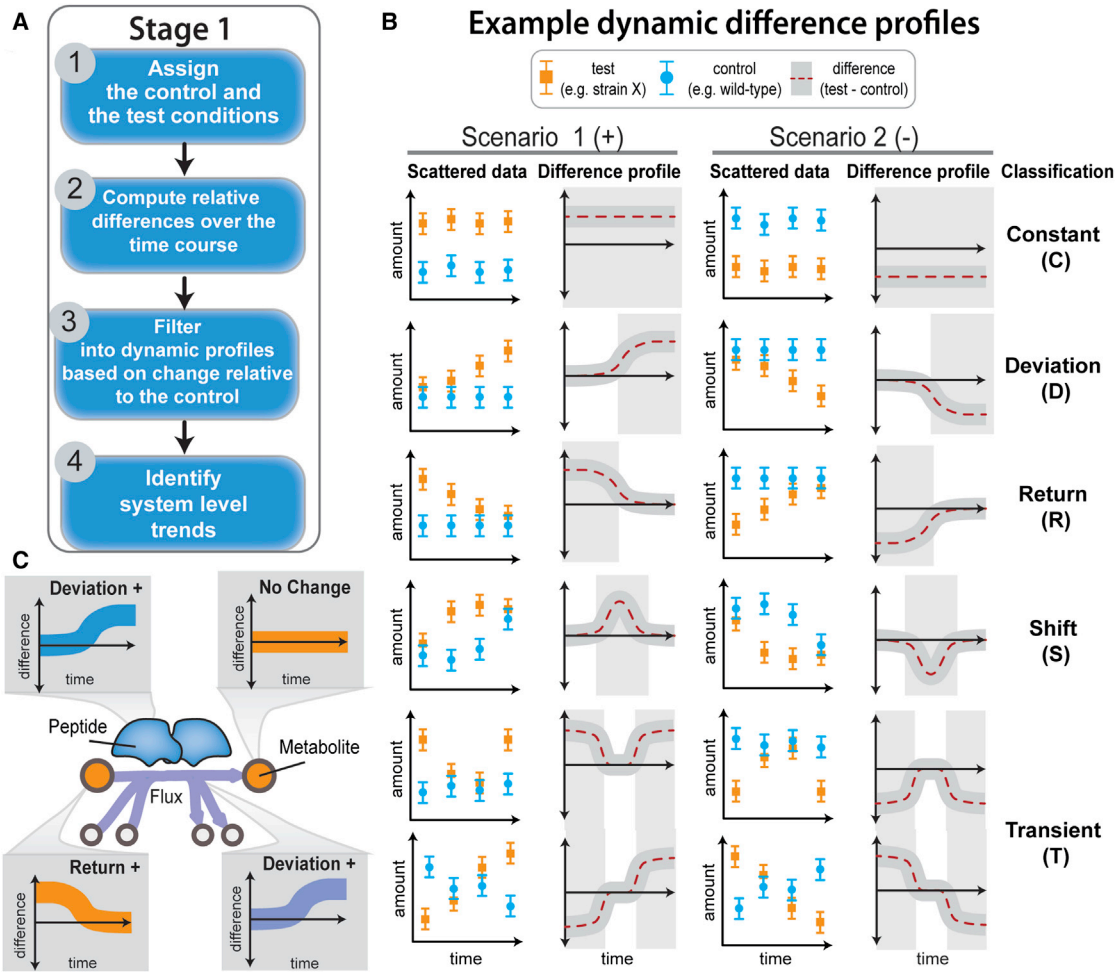


Figure 3. Systems-Level Multi-omics Integration and Analysis of Batch Fermentation Dynamics

(A) The first stage of the workflow filters, maps, and identifies system level differences between control (e.g., WT) and test (e.g., engineered strain) conditions through the construction of dynamic difference profiles.

(B) The differences for each data point relative to the control were calculated, and the errors of the measurements were propagated to determine the range of change (from significant to not changing) between the control and test conditions. The plots in the left column refer to positive ("+") shifts, or points where the test condition is greater than the control in terms of concentration or flux. Those in the right column refer to negative ("–") shifts. SDs for the test and control condition for each data point were calculated from triplicate measurements or estimated based on the percent root-squared deviation (%RSD) of representative triplicate measurements.

(C) A cartoon depiction of the data types included in this analysis: protein level measured by proteomics (top left), substrate and product metabolites measured through metabolomics (top right and bottom right), and computed flux (bottom right).

significantly shifted from WT levels. Strains I1, L1, and B1 consistently secrete acetate at similar levels to WT (e.g., "no change" or "constant" profiles; Figure S3), whereas strains I2, L2, and I3 strongly deviate (concentrations 14-, 15-, and 18-fold lower than WT). Dynamic difference profiles also highlight changes in less-abundant, intracellular ("_c") metabolites, where differences between strains are often more subtle. Certain strains show large-scale "transient" changes in the intracellular concentrations of citrate (cit_c), alpha-ketoglutarate (akg_c), glycolate (glyclt_c), and amino acids, such as glutamate and lysine, which are most dramatic for isopentenol producers—the strains that produce the most biofuel (Figure 2B).

Our findings generally point to a global pattern: the profiles of low-producing strains tend to "cluster" with WT rather than high-producing strains of the same fuel target. Despite the introduction of different heterologous pathways, the metabolite profiles of poorly optimized strains (i.e., I1, L1, and B1) show minimal deviations from WT. Similar to WT, these strains do not consume all available glucose and the concentrations of intracellular central carbon metabolites, like succinate (succ_c) and phosphoenolpyruvate (pep_c), match WT levels (e.g., "no change" profile). In contrast, profiles of top producing strains show large-scale deviations from WT, especially for citric acid cycle (TCA) metabolites (strains I2, I3, and L3 with 16- to 30-fold changes in concentration of succ_c and between 5- and 13-fold changes in concentration of pep_c for strains I2, I3, L2, and B2; Figure S3). These findings suggest that the level of pathway optimization, rather than the identity of the target

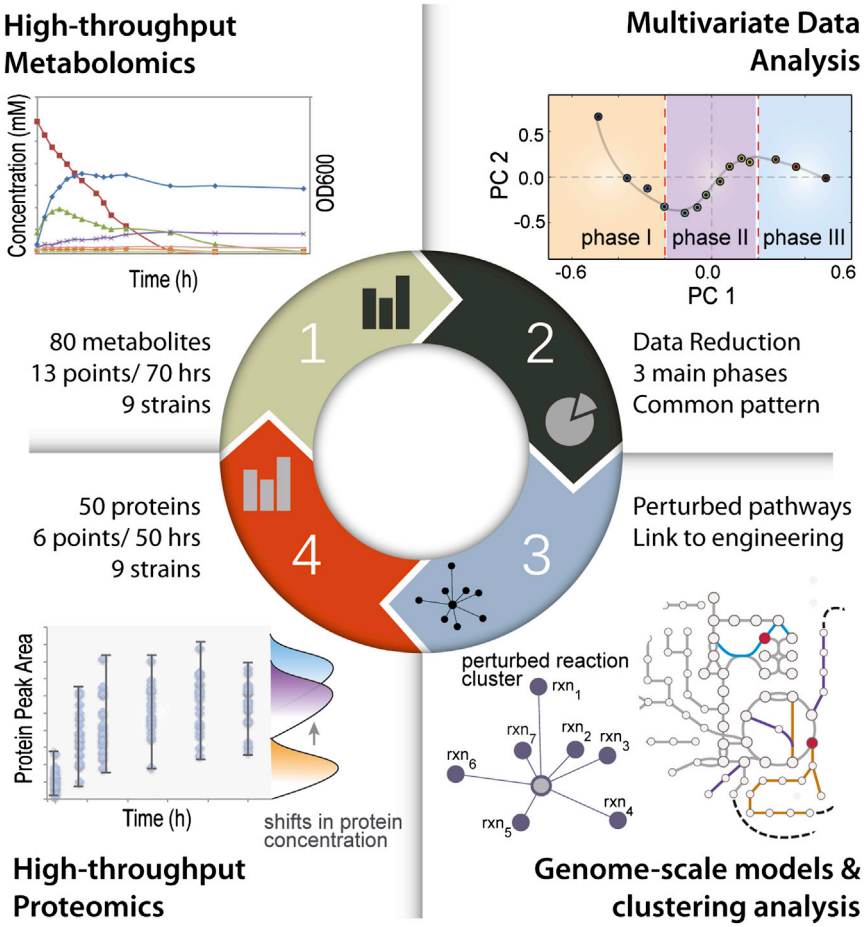


Figure 4. Integrating Multi-omics Data with Genome-Scale Models of Metabolism

Stages two and three of the workflow combine multivariate analyses and genome-scale models of metabolism. (1) Standard metabolomics and growth measurements for over 80 metabolites were taken for nine different strains over a 72-hr time course. (2) Applying PCA on this dataset, we find three distinct metabolic phases that align with different phases of the growth. (3) These pseudo-steady state phases are modeled using constraint-based methods, such as Markov-chain Monte Carlo-based sampling using extracellular measurements as inputs to the model. We compute and cluster perturbed reactions in host metabolism (illustrated by the red colored nodes in the network). (4) Perturbed reactions are assessed with other omics datasets, like proteomics.

biofuel, tends to dictate the endogenous metabolic response. While this is not entirely unexpected given the common mevalonate pathway "backbone" of each strain (Figure 2A), it suggests that the role of potentially confounding factors such as biofuel toxicity (Dunlop et al., 2011) or FPP (frdp_c) feedback inhibition (Primak et al., 2011), which vary markedly for each fuel target or pathway, is minimal in these strains compared to impact of overall product titer.

In summary, the first stage of this workflow provides a rapid means to filter complex omics data into categorical "dynamic difference profiles" and facilitate strain comparisons. The main understanding gained from this stage of the workflow is that, for the current group of strains, optimization level (i.e., overall product yield), rather than chosen fuel target, dictates the degree of metabolic perturbation. While this analysis provides valuable insight into the broad metabolic response to engineering by highlighting maximally perturbed nodes, additional analyses are needed to (1) explore how these changes are correlated over time (stage two) and (2) contextualize these perturbations within a biochemical network (stage three).

Stage Two: Correlations in Key Metabolic Fingerprints Distinguish Strain Behavior

Despite the high dimensionality of multi-omics datasets, unsupervised learning methods, such as Principal Component Analysis (PCA), capture much of the variation in a few key metabolites. In stage two of the workflow, we use standard multivariate analyses to reduce the dimensionality of multi-strain metabolomics data and identify common patterns in changing metabolite concentrations over time (Figure 4, steps 1 and 2). Specifically, we carry out PCA on the aggregate metabolomics dataset (nine strains, 13 time points, and 86 metabolites) to identify key metabolites that drive strain variation, determine how these drivers change over time, and uncover unique features for strain characterization.

Using PCA on this dataset shows that the first, second, and third singular vectors account for more than 80% of the variance in the dataset (Figure S4). For the top-producing strains, coefficients (factor loadings) for the fuel products tend to be the most significant, coinciding with the increased production yields for these strains. Not surprisingly, certain extracellular metabolites, including lactate, pyruvate, formate, and acetate, also have higher coefficients. Performing PCA on extracellular versus intracellular metabolites, we find that the first two eigenvectors sum to more than 60% and 70%, respectively, indicating that (1) changes in intracellular concentrations are correlated over time and (2) the uptake and secretion of extracellular metabolites are also correlated processes.

Plotting the first two singular vectors of PCA on the exometabolome shows a distinct three-state behavior in all nine strains. We find that these three phases correspond to distinct time intervals in the dataset: (1) phase I (0–6 hr); (2) phase II (6–20 hr); and (3) phase III (20–72 hr), as illustrated in a simplified depiction in Figure 4. The variation in each phase is driven by changes in extracellular metabolites, such as, in the case of WT, glucose in phase I, lactic acid, formate, and pyruvate in phase II, and acetate in phase III, which is consistent with what is commonly observed in exponential, early stationary, and late stationary growth phases of *E. coli*. These same metabolites show completely different behavior in top-producing strains (e.g.,

acetate becomes a driver of phase II in strains I3, L3, and B2 and formate and lactic acid drive phase III; Figure S5). The shift in acetate is interesting because its assimilation, or uptake, following its secretion is a key differentiator between optimized strains and non-optimized derivatives. By assimilating acetate in phase II, optimized strains such as I3 can recapture "lost" carbon and reform acetyl-CoA through the action of acetyl-CoA synthetase (Wolfe, 2005).

Intriguingly, changes in key intracellular metabolites also appear to coincide with this three phase behavior. As expected, amino acids are the main drivers of variation during the first phase. In the second phase, variation in low-producing strains is driven by glycolate (glyclt_c), glyoxylate (glx_c), and isocitrate (icit_c), which is consistent with glyoxylate metabolism and wild-type behavior. In top-producing strains, however, phosphoenolpyruvate (pep_c), citrate (cit_c), and α-ketoglutarate (akg_c) become the main drivers of phase II, which suggests the metabolic use of other TCA cycle reactions in these strains and corroborates the respective dynamic difference profiles from stage one.

To summarize, stage two of our workflow provides a means for correlating changes in metabolite concentrations over time. Using PCA, we identified three phases in time-course metabolomics data that are driven by the uptake and secretion of key metabolites, in addition to specific intracellular metabolite changes. The identification of these three metabolic phases motivates a more in-depth characterization of each of these states by genome scale modeling (stage three of our workflow). In the following section, we seek to understand whether the perturbed nodes discovered in the first stage of this workflow impact genome-scale flux networks. As described below, we use the findings from stage two to model pseudo-metabolic steady states.

Stage Three: Genome-Scale Modeling Provides Mechanistic Insights into Strain Variation

In stage three, genome-scale models provide contextual basis for the analysis of multi-scale omics datasets (O'Brien et al., 2015). Instead of only looking at one reaction, metabolite, or protein at a time, multiple reactions are modeled and assessed simultaneously, which helps in gaining insight into the reaction system as a whole. It is important to note that, while reduction of multidimensional data is an important principle of stage two, reduction of network-level information can be non-informative and misleading (e.g., if an important metabolic lead lies in a peripheral pathway not in core metabolism). Here, we use the comprehensive biochemical content of the metabolic network reconstruction of *E. coli* (Orth et al., 2011) and the predictive capability of constraint-based modeling approaches (O'Brien et al., 2015; Orth et al., 2010) to elucidate metabolic perturbations through the chemical connections contained in the reconstruction.

The identification of the three phases from PCA implies that each phase is a different metabolic state with a unique phenotype. While all nine strains share a similar characteristic three-phase behavior, the metabolites driving the variation in a given phase differ greatly (see stage two). This supports the hypothesis that even small variations in pathway engineering could lead to significant changes in endogenous metabolism. To investigate this, we carried out flux balance analysis (FBA) together with a Markov chain Monte Carlo-based (MCMC) sampling approach (Almaas et al., 2004) on each of the three phases for each strain. As discussed in detail below, this analysis shows that the exometabolome causes significant shifts in key reaction fluxes (p value <0.05 using an empirical test) relative to WT (Figure 4, steps 3 and 4). Most importantly, these shifting reactions cluster around the highly perturbed nodes that are observed in both metabolomic and proteomic datasets.

Significantly changing reaction fluxes indicate an increase or decrease in the flux (or flow of metabolites through a reaction), relative to WT. For each phase, we identified the most perturbed reaction fluxes, clustered the shifting reactions to find any common links between these nodes, and visualized the clusters graphically. The majority of shifting pathways include reactions in the pentose phosphate pathway (PPP), glycolysis/gluconeogenesis, and the TCA cycle (Figures 5A–5C, respectively), with the exception of some peripheral reactions (e.g., phosphopentomutase-2 deoxyribose). For high-producing strains, most of the significant shifting reactions (relative to WT) occur either in late exponential phase (phase I, Figure 5D) or early stationary phase (phase II, Figure 5E). Interestingly, we see strain-specific groupings in the types of reactions that shift significantly in these phases. For example, in strains L2 and B2 we observe increased flux through specific reactions in PPP, namely, phosphogluconate dehydrogenase (GND) and 6-phosphogluconolactonase (PGL), whereas for isopentenol-producing strains we see large-scale perturbation in flux networks surrounding triose phosphate isomerase (TPI), sedoheptulose 1,7-bisphosphate D-glyceraldehyde-3-phosphate-lyase (FBA3), transaldolase (TALA), and transketolase (TKT1, TKT2) (see Figures 5D and 5E). Other phenotypic changes become pronounced in certain strains during early stationary phase (phase II), such as flux diverting to TCA pathway reactions including alpha-ketoglutarate dehydrogenase (AKGDH), aconitase (ACONTa, ACONTb), citrate synthase (CS), and isocitrate dehydrogenase (ICDHyr) (Figures 5E and S6).

Visualization of these perturbed reaction nodes brings about a striking commonality that many are NADPH-producing reactions, such as ICDHyr, GND, and those related to specific amino acid biosynthetic pathways (e.g., AKGDH; Figure 6A). Ultimately, modeling indicates that the cumulative flux to NADPH-producing reactions is significantly elevated for higher-producing strains (Figure S6C). This observation is consistent with previous work that identified NADPH availability as a limiting factor in isoprenoid production in *Saccharomyces cerevisiae*, the native host for the mevalonate pathway (Asadollahi et al., 2009). One explanation for an apparent depletion of NADPH is related to the NADPH-dependent HMG-CoA reductase (HMGR), which catalyzes the second step of the heterologous mevalonate pathway in each engineered strain. Due to the action of this enzyme, two molecules of NADPH are consumed for each molecule of mevalonate that is produced, coupling high biofuel titers with increased demand for NADPH.

In summary, stage three of the workflow uses genome-scale modeling to elucidate the biological impacts of pathway optimization and highlight functional links between perturbed metabolic nodes. While the apparent NADPH limitation highlighted by these simulations seems obvious in retrospect, understanding how

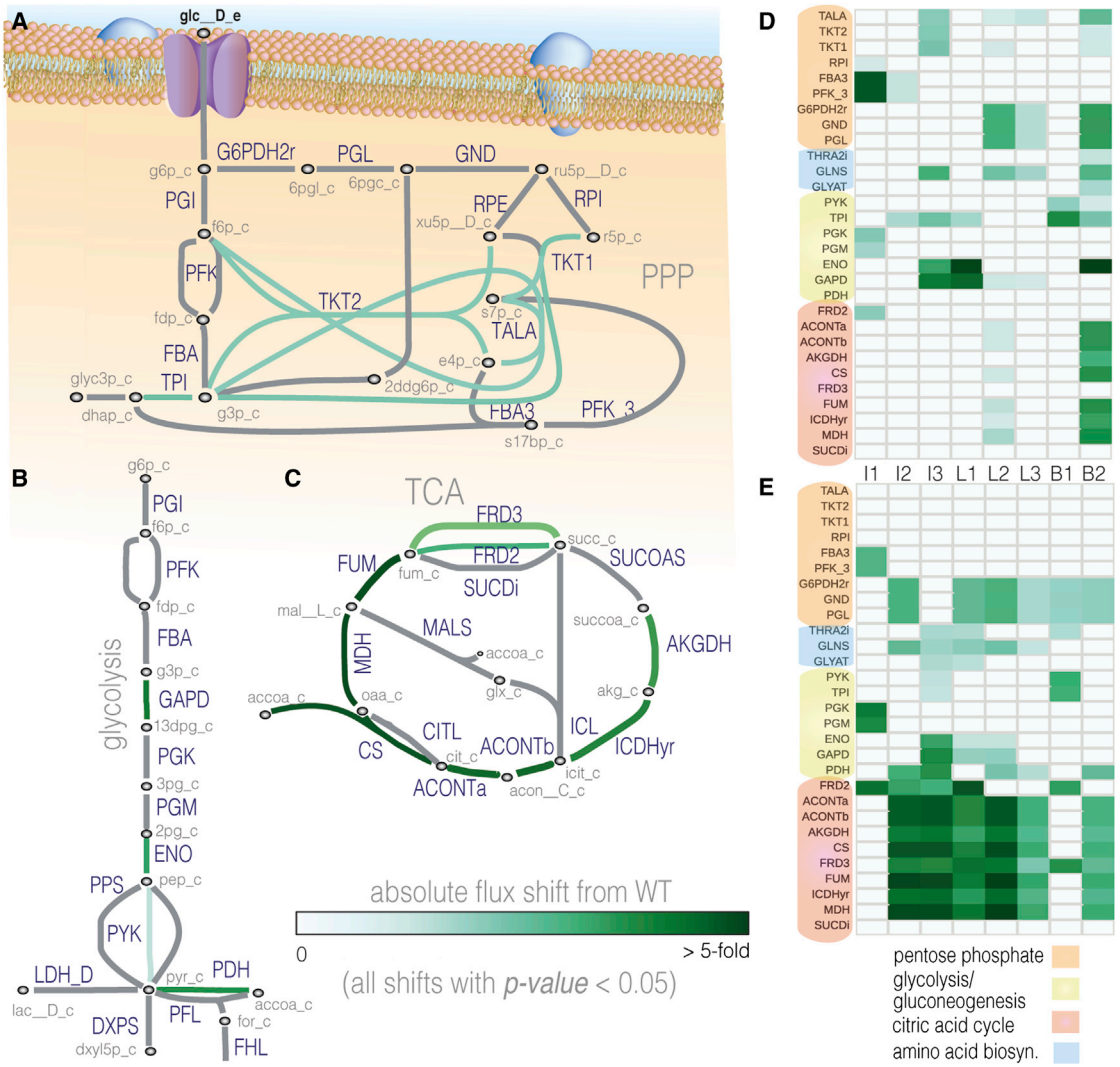


Figure 5. Genome-Scale Modeling Revealed Perturbations in TCA Cycle and Pentose-Phosphate Pathway Activity Associated with Certain Engineered Phenotypes

(A–C) Reactions colored by the shift (absolute value) in flux in a top-producing strain, I3, compared to wild-type in different pathways in central carbon metabolism: (A) the pentose-phosphate pathway; (B) glycolysis/gluconeogenesis; (C) TCA cycle.

(D and E) Shown in (D) are significant reaction flux shifts ($p < 0.05$) corresponding to various reactions in these pathways in phase I (0–6 hr) and those for phase II (6–20 hr) are displayed in (E). Here, shifts in metabolic flux represent overall changes (both positive and negative perturbations) from wild-type behavior.

networks re-route to accommodate such bottlenecks is less trivial. In the section that follows, we describe how tracing perturbations through a genome-scale flux network helps explain experimentally observed metabolic perturbations that, upon first glance, have no apparent connection with the NADPH node.

Model-Aided Predictions of Engineered Metabolic Phenotypes Are Consistent with Experiments

Perturbations in the intracellular flux networks, identified through modeling, can be cross-validated with complementary datasets, such as intracellular metabolomics and proteomics data. As mentioned in the above section, modeling intracellular flux networks makes use of uptake and secretion rates of glucose, organic acids, amino acids, and the fuel product. Therefore, the consistency of model predictions can be evaluated by comparing them to significantly perturbed nodes observed in the data. In this section, we demonstrate how three different data types, metabolomics, proteomics, and genome-scale flux predictions, are reconciled and corroborate our model-driven hypothesis that specific metabolic pathways are re-routed to meet the demands of pathway-induced NADPH depletion.

Our findings suggest that NADPH is depleted in engineered strains and that heterologous expression of HMGR is the main source of this behavior. The largest perturbations in reactions linked to a common NADPH node are found in strains expressing high levels of HMGR protein such as strain L2, which has HMGR levels 10- to 20-fold higher than any other strain. Consistent with increased flux through the HMGR reaction, intracellular concentrations of $NADP^+$ in strain L2 are significantly elevated compared to other strains (Figure 6B, box A), linking HMGR

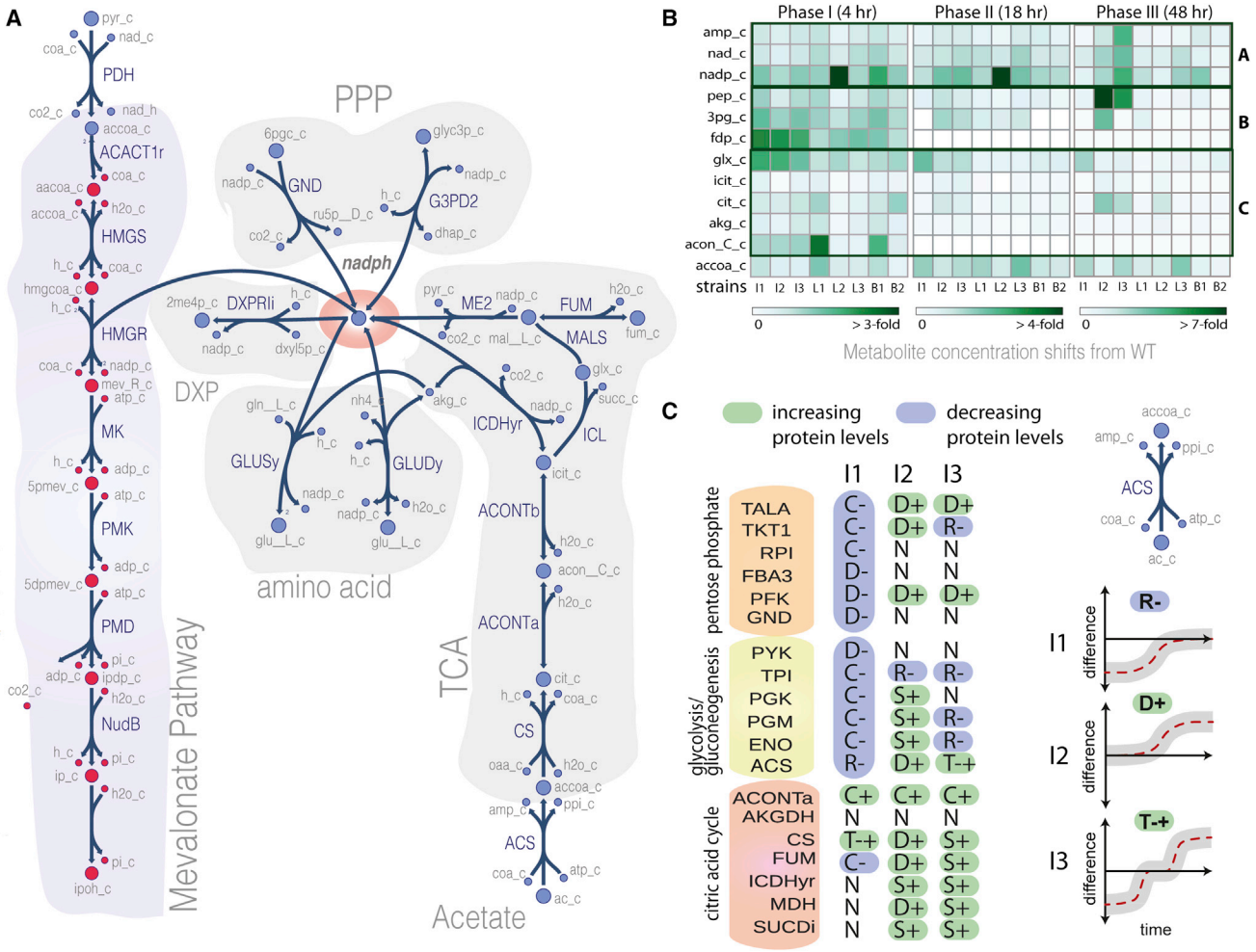


Figure 6. Constraint-Based Modeling Elucidates Pathways that Allow for Coupling of NADPH Metabolism and Biofuel Production

(A) The sum of flux through these main NADPH-producing and consuming reactions is significantly higher in top-producing strains over WT.

(B) Increases in cofactor (box A), glycolysis/gluconeogenesis (box B), and TCA (box C) metabolite concentrations (relative to wild-type *E. coli*) indicate regions in metabolism that are perturbed in different engineered strains.

(C) Dynamic difference profiles identify changes in protein levels for isopentenol-producing strains. As shown in the lower left panel, key glycolysis (yellow), PPP (orange), and TCA (red) proteins shift above WT levels in higher producing strains (I2 and I3). On the lower right panel is an example of how progressive engineering efforts change the dynamic difference profile for the protein acetate synthase (ACS).

expression with NADPH depletion. Furthermore, the cellular demand for NADPH appears to perturb several reactions in glycolysis and the TCA cycle in accordance with model predictions: phosphoglycerate, citrate, α-ketoglutarate, and malate levels increase by nearly 10-, 5-, and 3-fold, respectively, during the time course (see Figure 6B, boxes B and C).

While strain L2 is useful in establishing a clear link between HMGR expression and NADPH depletion, strain I3, the top performing strain on the basis of yield and product titer, provides even more convincing evidence of a metabolic response to pathway optimization, and, consequently, NADPH depletion. Model predictions for strain I3 indicate that key nodes in glycolysis/gluconeogenesis, such as glyceraldehyde-3-phosphate dehydrogenase (GAPD), triphosphate isomerase (TPI), and enolase (ENO), divert flux to provide the cell with routes to NADPH regeneration. One route for regenerating NADPH is through the PPP (e.g., GND and glucose-6-phosphate dehydrogenase, or G6PDH2r). Constructing dynamic difference profiles (stage one) from proteomics data, we find perturbations in key PPP proteins (e.g., G6PDH2r, GND, TALA, and TKT1) that are consistent with model predictions (Figure 6C; Table S2). Another route the cell uses for regenerating NADPH is through the TCA cycle (e.g., ICDHyr). Model predictions of increased flux through the TCA cycle are consistent with metabolomic measurements for strain I3: intracellular citrate and aconitase levels increase by 2- to 3-fold over WT (Figure 6B, box C) and dynamic difference profiles for proteins involved in these reactions increase by 2- to 3-fold over WT (e.g., CS and ACONTb protein levels; Figure 6C). In this context, a previously perplexing observation—the apparent "shunting" of assimilated acetate into the TCA cycle rather than the mevalonate pathway—is succinctly explained as a means to regenerate NADPH through ICDHyr rather than deplete it through HMGR.

A

Fitness characteristic	Gene knock-outs	Gene names
maximized flux through NADPH-producing reactions	b0197, b0198, b0199, b4238	metQ, metL, metN, nrdD
maximized flux through pentose-phosphate pathway	b1378, b0197, b0198, b0199, b4238	**ydbk**, metQ, metL, metN, nrdD
minimized flux through NADPH-consuming reactions	b4209, b1748, b1747, b2501, b4468	ytfE, astC, astA, ppk, glcE

B

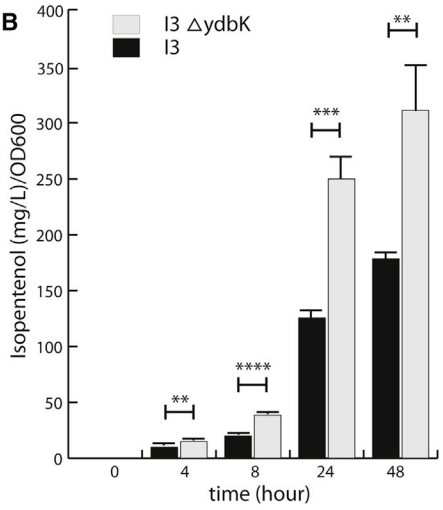


Figure 7. Model-Driven Predictions Discover a Gene Knockout that Increases the Specific Production of Isopentenol

(A) Single-gene knockout simulations were performed using genome-scale modeling to identify candidate targets that increase the production of isopentenol. Knockouts were experimentally tested and deletion of the gene in bold, *ydbK*, was found to increase specific production of isopentenol.

(B) Growth-normalized isopentenol titer (mg/l/OD 600) is displayed for strain I3 (black) and I3 with ΔydbK knockout (gray). At every non-zero time point, the knockout variant produces significantly more isopentenol than the highest producing strain, I3 (4 hr, **p = 0.0058; 8 hr, ****p < 0.0001; 24 hr, ***p = 0.002; 48 hr, **p = 0.0037, using an unpaired two-tailed t test). At 48 hr, absolute isopentenol titers are 920 mg/l versus 800 mg/l for strains I3ΔydbK and I3, respectively.

While we do not consistently observe significant shifts in the levels of some PPP proteins that would play an active role in NADPH regeneration, we do find significant increases (p value <0.05 using an empirical test) in RNA levels. GND expression, for instance, is increased by >5-fold over WT in strain 3A (unpublished data). Intriguingly, ENO and AKGDH expression levels are also 5-fold over WT, coinciding with increased levels of α-ketoglutarate-derived amino acids (e.g., glutamate, glutamine, histidine, and arginine) that were found to be enriched in high-producing isopentenol strains in stage one of this workflow.

Taken together, reconciliation of metabolomics and proteomics with genome-scale modeling proposes a general mechanism by which the cell responds to HMGR-mediated cofactor depletion by redirecting flux through the TCA cycle and/or PPP to regenerate NADPH. Importantly, the workflow highlights NADPH regeneration as a potential engineering strategy to improve mevalonate pathway function and product yields. As a validation, we attempted to address NADPH depletion not through mevalonate pathway engineering (Ma et al., 2011), but by identifying single gene knockouts (SKOs) that re-route flux to produce higher product yield.

Identifying Metabolic Properties Relevant to Re-engineering

Using the knowledge gained from the three-stage workflow, we reevaluated our constraint-based modeling simulations in the presence of single gene knockouts (SKOs). We were interested in discovering SKOs that re-route flux in pathways that compete with the mevalonate pathway and are related to NADPH production/depletion. As a proof of concept, we generated model-driven predictions of SKO candidates using the genome-scale metabolic model of strain I3, which produced the most biofuel and showed strong evidence for NADPH depletion.

Using flux variability analysis, we rank-ordered SKO candidates using several metrics that were identified to be important factors underlying strain variation in stage three of this workflow: (1) minimized flux through NADPH-consuming reactions; (2) maximized flux through NADPH-producing reactions; and (3) maximized flux through PPP reactions. The SKO candidates identified using these metrics are provided in Figure 7A and in the Supplemental Information (see Table S3 and an iPython notebook titled "Engineering_SKOs.ipynb").

To test the effects of the model-predicted SKOs, we experimentally tested the top four SKO candidates in strain I3 (Figure S7). We found that one of the SKOs, Δ*ydbK* (blattner code b1378), resulted in an almost 2-fold increase in the specific (i.e., growth-normalized) production of isopentenol (Figure 7B). In I3ΔydbK, specific production is increased at every time point during batch fermentation, and, by 48 hr, I3ΔydbK shows a higher absolute titer of isopentenol (920 mg/l versus 800 mg/l for strains I3ΔydbK and I3, respectively). This gene is predicted to act as a pyruvate synthase (reaction POR5 in the genome scale model, iJO1366 [Orth et. al. 2011]), which converts pyruvate to acetyl-CoA. Intriguingly, ΔydbK also significantly increases the specific production of limonene in strain L3, which also showed evidence of NADPH depletion (Figure S7), and suggests a commonality between isopentenol and limonene producing strains. In contrast, this SKO has minimal effects on bisabolene production in strain B2.

Conclusions

To date, the majority of metabolic engineering efforts serve as demonstrations of future potential rather than industry-ready technology. Achieving large-scale, economical production of microbial-derived products requires production strains to be exhaustively optimized for high yields and productivities. The challenges associated with this goal are numerous and massive in scope, including pathway "balancing" with appropriate protein expression and activity, product and metabolite toxicity,

feedback inhibition, strict energetic requirements, cofactor imbalances, and competition with endogenous pathways (Paddon and Keasling, 2014). Thus, the inherent complexity of biological systems makes them difficult to effectively design and control (Endy, 2005).

Greater synergy between systems biology, metabolic engineering, and synthetic biology would greatly benefit all three disciplines, given their complementary, yet classically separate, approaches to bioengineering (Nielsen et al., 2014). A major challenge in both metabolic engineering and synthetic biology is understanding how the introduction of engineered or non-native components into a biochemical network influences the behavior of the entire system. To meet this challenge, we have developed a three-stage workflow to interpret complex multi-omics data for multi-strain characterization. Each of the three stages of the workflow works together as a concerted pipeline to efficiently process highly dimensional datasets.

The first two stages of the workflow act as a flexible framework to interpret raw, multi-omics data by sorting strain phenotypes based on their dynamic difference profiles and correlating measurements based on distinct patterns derived from thousands of measurements. These two stages of the workflow in particular are well suited for integration with high-throughput strain engineering and analysis pipelines, where "manual" assessment of convoluted omics data is not feasible. While statistics-based approaches such as PCA can act as valuable stand-alone methods for metabolic engineering (Alonso-Gutierrez et al., 2015), unraveling the global response of the cell to pathway engineering requires moving beyond statistics-based approaches and incorporating system-wide analyses. To account for the systems-level response to pathway engineering, the third stage of this workflow leverages these statistics-based approaches in the context of a genome-scale metabolic model.

Here, we demonstrate that through the pairing of synthetic pathway construction and a systems-level, model-driven analysis, our multi-omics-based workflow successfully reconciles metabolomics data, proteomics data, and predictions from genome-scale models. Using mevalonate pathway engineering as a case study, we demonstrate that our approach is capable of elucidating the complex interplay between heterologous pathway engineering and endogenous metabolism in a microbial host. The utility of such a workflow is expected to become increasingly important with the parallel, accelerating advances in technologies related to strain generation and high-throughput analyses (e.g., on the order of thousands of strains and thousands of measurements).

EXPERIMENTAL PROCEDURES

All chemicals, media components, and enzymes were purchased from Sigma-Aldrich, VWR, or Fischer Scientific. *E. coli* DH10B and DH1 were purchased from Invitrogen and ATCC, respectively.

Plasmid and Strain Construction

Plasmids were assembled in *E. coli* DH10B according to the BglBrick standard (Anderson et al., 2010) using standardized vectors (Lee et al., 2011) with the exception of pTrc99A (Amann et al., 1988). Transformations into *E. coli* DH1 were performed with chemically competent cells as described previously (Chung et al., 1989). A list of plasmids and strains used in this study is provided in Table S1.

Growth Conditions and Production of Advanced Biofuels

Seed cultures of eight biofuel-producing strains (Table S1) and *E. coli* DH1 were grown overnight in 5-ml volumes of Luria broth (LB) medium with appropriate antibiotics at 37°C. For production, 100-ml volumes of EZ-Rich defined medium with 1% glucose in a 1-l Erlenmeyer flask were inoculated to an initial optical density 600 (OD_{600}) of 0.1, incubated with shaking (30°C, 200 rpm) to an OD_{600} of 0.6, and induced with 500 μM isopropyl β-D-1-thiogalactopyranoside (IPTG). For limonene- and bisabolene-producing strains, a 10% overlay of dodecane was added at induction. Mutant strains were grown in 25 ml of EZ-Rich defined media in 250-ml Erlenmeyer flasks with the same induction parameters.

Metabolomics and Proteomics Sampling and Analysis

Metabolomics samples (1.8 ml) were collected throughout the fermentation at set time points. For isopentenol strains, 0.2 ml of culture was extracted with ethyl acetate for gas chromatography-flame ionization detector (GC-FID) analysis (George et al., 2014). For limonene- and bisabolene-producing strains, 0.1 ml of dodecane overlay was collected and diluted into ethyl acetate for analysis by gas chromatography-mass spectrometry (GC-MS) (Alonso-Gutierrez et al., 2013; Peralta-Yahya et al., 2011). After removing 0.1 ml to measure OD600, the remaining volume (1.5 ml) was pelleted (14,000 × *g*) in a tabletop centrifuge at 4°C. Supernatant (0.25 ml) was collected for organic acid analysis by high performance liquid chromatography (HPLC), and another 0.25 ml of supernatant was mixed 1:1 with ice cold MeOH and stored at −20°C for the quantification of extracellular metabolites. The remaining supernatant was decanted, the pellet was resuspended in 0.3 ml of ice-cold MeOH by vortexing, and the suspension was centrifuged (8000 × *g* at 4°C) for 10 min. The supernatant was collected, mixed with a 1:1 volume of water, and filtered through a Millipore Amicon Ultra 3 kDa molecular weight (MW) cutoff filter (14,000 × *g* for 60 min at −2°C). Water was added to the flow-through to a final volume of 1 ml, and the samples were lyophilized overnight. Samples were reconstituted in 90 μl MeOH:H_2O (1:1) prior to analysis.

Organic acids were analyzed by an Agilent 1200 Series HPLC system equipped with a photodiode array detector set at 210, 254, and 280 nm. Metabolites were isocratically separated (4 mM sulfuric acid, flow rate of 0.6 ml/min) on an Aminex HPX-87H column with 8% cross linkage (150 mm length, 7.8 mm internal diameter, 9 μm particle size; Bio-Rad). Intracellular and extracellular metabolites were analyzed by liquid chromatography mass spectrometry (LC-MS) on a ZIC-HILIC column (150 mm length, 2.1 mm internal diameter, 2.5 μm particle size) using an Agilent 1200 Series HPLC coupled to an Agilent 6210 time-of-flight mass spectrometer. Metabolites were quantified via eight-point calibration curves ranging from 781.25 nM to 200 μM. Please see previous references for details on the quantification of glycolysis and TCA cycle intermediates (Juminaga et al., 2012), amino acids (Bokinsky et al., 2013), organic acids (Juminaga et al., 2012), nucleotides and CoAs (Bokinsky et al., 2013), and mevalonate pathway intermediates (Weaver et al., 2015).

Proteomics samples (5 ml) were harvested by centrifugation in a 15-ml falcon tube (5,000 × *g* at 4°C), supernatant was decanted, and cell pellets were stored at −80°C. Pellets were extracted with chloroform/methanol and protein samples were prepared as described previously (Redding-Johanson et al., 2011). Following drying in a vacuum concentrator (ThermoSavant), the protein pellet was resuspended in ammonium bicarbonate and quantified using DC Protein reagent (Bio-Rad). Protein was diluted to 0.5 mg/ml, and disulfide bonds were reduced with tris(2-carboxyethyl)phosphine (TCEP) for 30 min at room temperature followed by disulfide bond blocking with 10 mM iodoacetamide at room temperature in the dark for 30 min. Samples were analyzed using an AB Sciex 5500Q-Trap mass spectrometer operating in MRM (SRM) mode coupled to an Agilent 1100 system. For method details, please see references George et al. (2014) and Batth et al. (2014).

Constraint-Based Modeling

Constraint-based modeling and analysis of metabolic networks have been extensively reviewed and described elsewhere (Bordbar et al., 2014). Once the topology and set of constraints is known, the model can be used with various constraint-based methods to understand or predict cellular phenotypes.

We started with the *i*JO1366 model of *E. coli* K12-MG1655 (Orth et al., 2011). Heterologous genes and non-native mevalonate pathway intermediates were

added to the model in the form of mass and charge balanced reactions. Select metabolites, known to cross the cell membrane (based on extracellular measurements), were added as exchange reactions. When available, uptake and secretion rates were used from new or published data to constrain the upper and lower bounds of the exchange reactions (George et al., 2014). All parameters are detailed in Tables S1 and S2. Markov chain Monte Carlo (MCMC) sampling was used to generate a set of feasible distributions of fluxes in the genome-scale network. If carried out for enough time, the set of points will distribute uniformly throughout flux space and will provide a distribution (or range) in fluxes through a given reaction. This range represents most likely flux for a given reaction in the metabolic network, and depends on the network topology and model constraints. A *Z*-score-based analysis was carried out to determine the most significantly shifting fluxes between wild-type and engineered strains as well as between various regions of growth (i.e., exponential versus stationary phase). The *Z*-score calculation was repeated 1,000 times, and the mean value is reported. Comparisons with the experimental data were done by calculating differences in concentration and peak areas for the metabolomics and proteomics datasets. Similarly, *Z*-scores were computed to determine which shifts were significant over others.

iPython Notebooks

The three-stage iPython notebook series files can be found at the following links. Stage one: https://github.com/SBRG/Strain_characterization_workflow/blob/master/ipython_notebook/Stage_one_Dynamic_Differences.ipynb; stage two: https://github.com/SBRG/Strain_characterization_workflow/blob/master/ipython_notebook/Stage_two_multivariate.ipynb; and stage three: https://github.com/SBRG/Strain_characterization_workflow/blob/master/ipython_notebook/Stage_three_GEM.ipynb.

iPython Notebook for SKO Engineering

iPython notebook for SKO engineering can be found at the following link: https://github.com/SBRG/Strain_characterization_workflow/blob/master/ipython_notebook/Engineering_SKOs.ipynb.

SUPPLEMENTAL INFORMATION

Supplemental Information includes Supplemental Experimental Procedures, seven figures, three tables, and two data files and can be found with this article online at http://dx.doi.org/10.1016/j.cels.2016.04.004.

AUTHOR CONTRIBUTIONS

Conceptualization, K.W.G., E.B., J.A.-G., and T.S.L.; Methodology, E.B., K.W.G., J.A.-G., C.J.P., T.S.B., L.Y., D.M., and J.M.; Investigation, E.B., K.W.G, J.A-G., and M.T.; Writing – Original Draft, E.B. and K.W.G.; Writing – Review & Editing, E.B., K.W.G., T.S.L., B.O.P., J.D.K., P.D.A., H.G.M., A.F., E.J.O., and C.J.P.; Funding Acquisition, T.S.L and J.D.K; Resources, T.S.L. and J.D.K.; Supervision, T.S.L, J.D.K., and B.O.P.

ACKNOWLEDGMENTS

This work was funded by the Joint BioEnergy Institute (http://www.jbei.org/), which is supported by the US Department of Energy, Office of Science, Office of Biological and Environmental Research, through contract DE-AC02-05CH11231 between Lawrence Berkeley National Laboratory and the US Department of Energy (to K.W.G., J.A.-G., M.T., H.G.M., E.B., C.J.P., P.D.A., J.D.K., and T.S.L.); the Swiss National Science Foundation (grant p2elp2_148961 to E.B.); and the National Institutes of Health (grant GM057089 to B.O.P.). This research used resources of the National Energy Research Scientific Computing Center, a DOE Office of Science User Facility supported by the Office of Science of the U.S. Department of Energy under Contract No. DE-AC02-05CH11231. We also gratefully acknowledge Dr. Daniel Zielinski, Dr. Aarash Bordbar, Chris Shymansky, Jennifer Gin, Dr. Josh Lerman, and Professor Vassily Hatzimanikatis for early input in the project as well as Ali Ebrahim for technical support. The United States Government retains, and the publisher, by accepting the article for publication, acknowledges that the United States Government retains a non-exclusive, paid-up, irrevocable, world-wide license to publish or reproduce the published form of this manuscript, or allow others to do so, for United States Government purposes.

Received: September 30, 2015
Revised: February 18, 2016
Accepted: April 4, 2016
Published: May 19, 2016

REFERENCES

Almaas, E., Kovács, B., Vicsek, T., Oltvai, Z.N., and Barabási, A.-L. (2004). Global organization of metabolic fluxes in the bacterium Escherichia coli. Nature *427*, 839–843.

Alonso-Gutierrez, J., Chan, R., Batth, T.S., Adams, P.D., Keasling, J.D., Petzold, C.J., and Lee, T.S. (2013). Metabolic engineering of Escherichia coli for limonene and perillyl alcohol production. Metab. Eng. *19*, 33–41.

Alonso-Gutierrez, J., Kim, E.-M., Batth, T.S., Cho, N., Hu, Q., Chan, L.J.G., Petzold, C.J., Hillson, N.J., Adams, P.D., Keasling, J.D., et al. (2015). Principal component analysis of proteomics (PCAP) as a tool to direct metabolic engineering. Metab. Eng. *28*, 123–133.

Amann, E., Ochs, B., and Abel, K.J. (1988). Tightly regulated tac promoter vectors useful for the expression of unfused and fused proteins in Escherichia coli. Gene *69*, 301–315.

Anderson, J.C., Dueber, J.E., Leguia, M., Wu, G.C., Goler, J.A., Arkin, A.P., and Keasling, J.D. (2010). BglBricks: A flexible standard for biological part assembly. J. Biol. Eng. *4*, 1.

Asadollahi, M.A., Maury, J., Patil, K.R., Schalk, M., Clark, A., and Nielsen, J. (2009). Enhancing sesquiterpene production in Saccharomyces cerevisiae through in silico driven metabolic engineering. Metab. Eng. *11*, 328–334.

Bailey, J.E. (1991). Toward a science of metabolic engineering. Science *252*, 1668–1675.

Batth, T.S., Singh, P., Ramakrishnan, V.R., Sousa, M.M.L., Chan, L.J.G., Tran, H.M., Luning, E.G., Pan, E.H.Y., Vuu, K.M., Keasling, J.D., et al. (2014). A targeted proteomics toolkit for high-throughput absolute quantification of Escherichia coli proteins. Metab. Eng. *26*, 48–56.

Berger, B., Peng, J., and Singh, M. (2013). Computational solutions for omics data. Nat. Rev. Genet. *14*, 333–346.

Bokinsky, G., Baidoo, E.E.K., Akella, S., Burd, H., Weaver, D., Alonso-Gutierrez, J., García-Martín, H., Lee, T.S., and Keasling, J.D. (2013). HipA-triggered growth arrest and β-lactam tolerance in Escherichia coli are mediated by RelA-dependent ppGpp synthesis. J. Bacteriol. *195*, 3173–3182.

Bordbar, A., Monk, J.M., King, Z.A., and Palsson, B.O. (2014). Constraint-based models predict metabolic and associated cellular functions. Nat. Rev. Genet. *15*, 107–120.

Chou, H.H., and Keasling, J.D. (2012). Synthetic pathway for production of five-carbon alcohols from isopentenyl diphosphate. Appl. Environ. Microbiol. *78*, 7849–7855.

Chung, C.T., Niemela, S.L., and Miller, R.H. (1989). One-step preparation of competent Escherichia coli: transformation and storage of bacterial cells in the same solution. Proc. Natl. Acad. Sci. USA *86*, 2172–2175.

de Jong, B., Siewers, V., and Nielsen, J. (2012). Systems biology of yeast: enabling technology for development of cell factories for production of advanced biofuels. Curr. Opin. Biotechnol. *23*, 624–630.

Dunlop, M.J., Dossani, Z.Y., Szmidt, H.L., Chu, H.C., Lee, T.S., Keasling, J.D., Hadi, M.Z., and Mukhopadhyay, A. (2011). Engineering microbial biofuel tolerance and export using efflux pumps. Mol. Syst. Biol. *7*, 487.

Endy, D. (2005). Foundations for engineering biology. Nature *438*, 449–453.

Fuhrer, T., and Zamboni, N. (2015). High-throughput discovery metabolomics. Curr. Opin. Biotechnol. *31*, 73–78.

George, K.W., Chen, A., Jain, A., Batth, T.S., Baidoo, E.E.K., Wang, G., Adams, P.D., Petzold, C.J., Keasling, J.D., and Lee, T.S. (2014). Correlation analysis of targeted proteins and metabolites to assess and engineer microbial isopentenol production. Biotechnol. Bioeng. *111*, 1648–1658.

George, K.W., Alonso-Gutierrez, J., Keasling, J.D., and Lee, T.S. (2015a). Isoprenoid drugs, biofuels, and chemicals—artemisinin, farnesene, and beyond. Adv. Biochem. Eng. Biotechnol. *148*, 355–389.

George, K.W., Thompson, M.G., Kang, A., Baidoo, E., Wang, G., Chan, L.J.G., Adams, P.D., Petzold, C.J., Keasling, J.D., and Lee, T.S. (2015b). Metabolic engineering for the high-yield production of isoprenoid-based C_5 alcohols in E. coli. Sci. Rep. *5*, 11128.

Gross, M. (2011). Riding the wave of biological data. Curr. Biol. *21*, R204–R206.

Han, M.J., Yoon, S.S., and Lee, S.Y. (2001). Proteome analysis of metabolically engineered Escherichia coli producing Poly(3-hydroxybutyrate). J. Bacteriol. *183*, 301–308.

Han, M.-J., Jeong, K.J., Yoo, J.-S., and Lee, S.Y. (2003). Engineering Escherichia coli for increased productivity of serine-rich proteins based on proteome profiling. Appl. Environ. Microbiol. *69*, 5772–5781.

Jullesson, D., David, F., Pfleger, B., and Nielsen, J. (2015). Impact of synthetic biology and metabolic engineering on industrial production of fine chemicals. Biotechnol. Adv. *33*, 1395–1402.

Juminaga, D., Baidoo, E.E.K., Redding-Johanson, A.M., Batth, T.S., Burd, H., Mukhopadhyay, A., Petzold, C.J., and Keasling, J.D. (2012). Modular engineering of L-tyrosine production in Escherichia coli. Appl. Environ. Microbiol. *78*, 89–98.

Kabir, M.M., and Shimizu, K. (2003). Fermentation characteristics and protein expression patterns in a recombinant Escherichia coli mutant lacking phosphoglucose isomerase for poly(3-hydroxybutyrate) production. Appl. Microbiol. Biotechnol. *62*, 244–255.

Kahn, S.D. (2011). On the future of genomic data. Science *331*, 728–729.

Kuehnbaum, N.L., and Britz-McKibbin, P. (2013). New advances in separation science for metabolomics: resolving chemical diversity in a post-genomic era. Chem. Rev. *113*, 2437–2468.

Kwok, R. (2010). Five hard truths for synthetic biology. Nature *463*, 288–290.

Landels, A., Evans, C., Noirel, J., and Wright, P.C. (2015). Advances in proteomics for production strain analysis. Curr. Opin. Biotechnol. *35*, 111–117.

Lee, S.Y., Lee, D.-Y., and Kim, T.Y. (2005). Systems biotechnology for strain improvement. Trends Biotechnol. *23*, 349–358.

Lee, T.S., Krupa, R.A., Zhang, F., Hajimorad, M., Holtz, W.J., Prasad, N., Lee, S.K., and Keasling, J.D. (2011). BglBrick vectors and datasheets: A synthetic biology platform for gene expression. J. Biol. Eng. *5*, 12.

Ma, S.M., Garcia, D.E., Redding-Johanson, A.M., Friedland, G.D., Chan, R., Batth, T.S., Haliburton, J.R., Chivian, D., Keasling, J.D., Petzold, C.J., et al. (2011). Optimization of a heterologous mevalonate pathway through the use of variant HMG-CoA reductases. Metab. Eng. *13*, 588–597.

Margolis, R., Derr, L., Dunn, M., Huerta, M., Larkin, J., Sheehan, J., Guyer, M., and Green, E.D. (2014). The National Institutes of Health's Big Data to Knowledge (BD2K) initiative: capitalizing on biomedical big data. J. Am. Med. Inform. Assoc. *21*, 957–958.

Martin, V.J.J., Pitera, D.J., Withers, S.T., Newman, J.D., and Keasling, J.D. (2003). Engineering a mevalonate pathway in Escherichia coli for production of terpenoids. Nat. Biotechnol. *21*, 796–802.

Metzker, M.L. (2010). Sequencing technologies - the next generation. Nat. Rev. Genet. *11*, 31–46.

Nielsen, J., Fussenegger, M., Keasling, J., Lee, S.Y., Liao, J.C., Prather, K., and Palsson, B. (2014). Engineering synergy in biotechnology. Nat. Chem. Biol. *10*, 319–322.

O'Brien, E.J., Monk, J.M., and Palsson, B.O. (2015). Using genome-scale models to predict biological capabilities. Cell *161*, 971–987.

Orth, J.D., Thiele, I., and Palsson, B.Ø. (2010). What is flux balance analysis? Nat. Biotechnol. *28*, 245–248.

Orth, J.D., Conrad, T.M., Na, J., Lerman, J.A., Nam, H., Feist, A.M., and Palsson, B.Ø. (2011). A comprehensive genome-scale reconstruction of Escherichia coli metabolism–2011. Mol. Syst. Biol. *7*, 535.

Paddon, C.J., and Keasling, J.D. (2014). Semi-synthetic artemisinin: a model for the use of synthetic biology in pharmaceutical development. Nat. Rev. Microbiol. *12*, 355–367.

Palsson, B., and Zengler, K. (2010). The challenges of integrating multi-omic data sets. Nat. Chem. Biol. *6*, 787–789.

Peralta-Yahya, P.P., Ouellet, M., Chan, R., Mukhopadhyay, A., Keasling, J.D., and Lee, T.S. (2011). Identification and microbial production of a terpene-based advanced biofuel. Nat. Commun. *2*, 483.

Peralta-Yahya, P.P., Zhang, F., del Cardayre, S.B., and Keasling, J.D. (2012). Microbial engineering for the production of advanced biofuels. Nature *488*, 320–328.

Picotti, P., and Aebersold, R. (2012). Selected reaction monitoring-based proteomics: workflows, potential, pitfalls and future directions. Nat. Methods *9*, 555–566.

Primak, Y.A., Du, M., Miller, M.C., Wells, D.H., Nielsen, A.T., Weyler, W., and Beck, Z.Q. (2011). Characterization of a feedback-resistant mevalonate kinase from the archaeon Methanosarcina mazei. Appl. Environ. Microbiol. *77*, 7772–7778.

Redding-Johanson, A.M., Batth, T.S., Chan, R., Krupa, R., Szmidt, H.L., Adams, P.D., Keasling, J.D., Lee, T.S., Mukhopadhyay, A., and Petzold, C.J. (2011). Targeted proteomics for metabolic pathway optimization: application to terpene production. Metab. Eng. *13*, 194–203.

Rolfsson, Ó., and Palsson, B.O. (2015). Decoding the jargon of bottom-up metabolic systems biology. BioEssays *37*, 588–591.

Stephens, Z.D., Lee, S.Y., Faghri, F., Campbell, R.H., Zhai, C., Efron, M.J., Iyer, R., Schatz, M.C., Sinha, S., and Robinson, G.E. (2015). Big data: astronomical or genomical? PLoS Biol. *13*, e1002195.

Tyo, K.E., Alper, H.S., and Stephanopoulos, G.N. (2007). Expanding the metabolic engineering toolbox: more options to engineer cells. Trends Biotechnol. *25*, 132–137.

Weaver, L.J., Sousa, M.M.L., Wang, G., Baidoo, E., Petzold, C.J., and Keasling, J.D. (2015). A kinetic-based approach to understanding heterologous mevalonate pathway function in E. coli. Biotechnol. Bioeng. *112*, 111–119.

Wolfe, A.J. (2005). The acetate switch. Microbiol. Mol. Biol. Rev. *69*, 12–50.

Zhang, Z., Wu, S., Stenoien, D.L., and Paša-Tolić, L. (2014). High-throughput proteomics. Annu. Rev. Anal. Chem. *7*, 427–454.